Mineração de Dados: Definição, Técnicas e Aplicabilidade

Renan Ruseler

Fundação Universidade do Estado de Santa Catarina (UDESC – Ceavi)

E-mail: renan.ruseler@gmail.com

Resumo: Este artigo apresenta uma revisão acerca da Mineração de Dados como a procura de padrões consistentes através da exploração de grandes quantidades de dados. Consiste, basicamente, na descoberta de padrões interessantes e potencialmente úteis de grandes conjuntos de dados e algoritmos para a extração de informações ocultas. Ela aborda áreas de pesquisa que envolvem estatística, matemática e computação, podendo ser aplicada tanto para pesquisas científicas quanto para impulsionar a lucratividade das empresas. Dessa maneira, a Mineração de Dados está se tornando cada vez mais popular como uma ferramenta de descoberta de informações valiosas que possam orientar decisões em condições de certeza limitada. O conhecimento adquirido auxilia as organizações principalmente no desenvolvimento de estratégias e na gestão dos negócios. Trata-se de um tópico recente da ciência da computação que utiliza inúmeras técnicas, modelos e algoritmos de áreas como a de reconhecimento de padrões, estatística e inteligência artificial.

Palavras-chave: mineração, dados, algoritmos, informações.

Abstract: *This article presents a review of Data Mining as the search for consistent patterns by exploiting large amounts of data. It basically consists in discovering interesting and potentially useful patterns of large data sets and algorithms for extracting hidden information. It adresses research areas involving statistics, mathematics and computing, and can be applied to both scientific reseach and business profitability. In this way, Data Mining is becoming increasingly popular as a tool for discovering valuable information that can guide decisions in conditions of limited certainty. The acquired knowledge assists the organizations mainly in the development of strategies and in the managment of the business. This is a recent topic of computer science that uses innumerable techniques, models and algorithms from several areas such as pattern recognition, statistics and artificial intelligence.*

Keywords: *mining, datas, algorithms, informations.*

1. Introdução

No atual universo corporativo, devido à altíssima competitividade dos mercados, é necessário que as empresas invistam em sistemas cada vez mais complexos e sofisticados capazes de auxiliá-los na dinâmica do seu negócio. Tais sistemas alimentam bases de dados que, por representarem a realidade cotidiana do negócio daquela empresa e por possuírem informações valiosas, são de extrema importância para a organização. O que acontece muitas vezes é que, ao se depararem com um volume monstruoso de dados, muitas pessoas não conseguem interpretá-los e decifrá-los da maneira correta, pois isso requer um nível de capacidade técnica e analítica

muito grande, o que pode ser um problema até mesmo para profissionais e especialistas da área. Nesse contexto, faz-se necessário a utilização de algoritmos e ferramentas que possam extrair dessas bases de dados, de maneira rápida e segura, informações precisas e consistentes para que os usuários possam utilizá-las visando o benefício empresarial e o desenvolvimento de estratégias e planejamentos a curto, médio e a longo prazo. Tais informações podem ser de imensa ajuda no desenvolvimento da organização e nas tomadas de decisão. É nesse encadeamento de ideias que as ferramentas, métodos e técnicas da Mineração de Dados se aplicam.

Assim sendo, a Mineração de Dados está entre as tecnologias mais promissoras da atualidade. Ela possui o poder de transformar tabelas de números em palavras e traduzir dados brutos em linguagem empresarial. Através dela, empresas podem descobrir tendências e padrões de comportamento que antes passavam despercebidos. Por evoluir muito no decorrer dos anos e por ser considerada uma área de pesquisa multidisciplinar, abrangendo setores como o da segurança, da medicina e até mesmo do telemarketing, as definições acerca do seu termo variam muito de acordo com a época e com a área de atuação dos autores. Este estudo tem como objetivo apresentar uma visão detalhada acerca da Mineração de Dados. Ao longo deste artigo, são apresentadas as definições, os aspectos, suas principais técnicas, as funcionalidades e a importância da Mineração de Dados no atual cenário global. A seção 2 apresenta as definições de o que seria a Mineração de Dados. A seção 3 descreve as principais técnicas e metodologias que a compõem. Na seção 4, encontra-se uma visão geral acerca das suas utilidades, suas funcionalidades, sua aplicabilidade e da sua importância no mundo dos negócios, principalmente nas grandes empresas e corporações. Ao final deste estudo, na seção 5, encontram-se as considerações finais.

2. Definição

O conceito de Mineração de Dados (*Data Mining*) tem se tornado cada vez mais popular como sendo o processo de explorar grandes quantidades de dados à procura de padrões consistentes para detectar relacionamentos sistemáticos entre variáveis e, consequentemente, novos subconjuntos de dados. Conforme explica Carvalho, pode-se definir está área como “[...] o uso de técnicas automáticas de exploração de grandes quantidades de dados de forma a descobrir novos padrões e relações que, devido ao volume de dados, não seriam facilmente descobertos a olho nu pelo ser humano” (2001, p. 7). É considerada uma área de pesquisa multidisciplinar, incluindo principalmente as tecnologias de banco de dados, reconhecimento de padrões, teoria da probabilidade, inteligência artificial, estatística e visualização de dados. Ela é uma maneira de procurar relações interessantes escondidas em um grande conjunto de dados, tais como padrões de agrupamentos e aproximações de funções. Também pode ser definida como uma metodologia que procura uma descrição lógica ou matemática, possivelmente de natureza complexa, de padrões que representam relações sistemáticas entre variáveis em um conjunto de dados. Ela procura obter, a partir dos dados operativos brutos, informações úteis para subsidiar as tomadas de decisão.

Devido ao grande montante de dados, esses padrões dificilmente seriam descobertos através de métodos tradicionais como consultas a base de dados ou relatórios. Através dos resultados obtidos, é possível desenvolver pesquisas científicas

e elaborar estratégias para impulsionar a lucratividade de uma empresa, tornando-a inovadora e competitiva no mercado. A técnica de mineração almeja principalmente a realização de previsões com possíveis fatos e correlações não explicitadas. Sua funcionalidade irá especificar que tipo de padrões ou relacionamentos entre os registros e suas variáveis podem ser utilizados na mineração. O processo de mineração pode ser aplicado a qualquer massa de dados, sejam eles oriundos da astronomia, medicina, geologia, economia, administração, engenharia, entre outras áreas de estudo. De maneira simplificada, ele pode ser dividido nas seguintes etapas: exploração, construção de modelo, definição de padrão, validação e verificação. Tal processo não se resume em analisar grandes quantidades de massas de informação e descobrir novas relações. É preciso atuar sobre essas relações de forma a transformar dados brutos em conhecimento e, consequentemente, em algum tipo de retorno empresarial.

A Mineração de Dados é considerada “[...] uma metodologia auxiliar, que otimiza o processo de escolha de descritores e limiares das classes, pois procura definir padrões, associações, mudanças, anomalias e estruturas significativas entre os dados” (NASCIMENTO; CRUZ, p. 2346). Ela possui uma vasta aplicação nos mais diferentes segmentos, seja na parte governamental, corporativa ou acadêmica, além de uma série de desafios relevantes que podem dar início a excelentes trabalhos científicos. Existem dois conceitos muito importantes que devem ser abordados a respeito dessa tecnologia, que são: funcionalidades e técnicas. As funcionalidades são aplicações, funções ou utilidades que especificam o que se quer buscar no conjunto de dados. As técnicas, por sua vez, são estratégias, processos ou métodos utilizados para descobrir os padrões de interesse da pesquisa. Algumas aplicações trafegam um volume altíssimo de dados, fazendo com que o processo de mineração se torne uma tarefa inviável. Frequentemente, são necessárias técnicas que lidem com bases de dados cada vez maiores, chegando na casa dos *Terabytes*. Esse processo costuma ocorrer utilizando dados contidos dentro do *Data Warehouse*, pois ele armazena e apresenta grandes volumes de dados de maneira mais consistente, sistemática e organizada.

3. Técnicas

Antes de mais nada, é necessário deixar claro que não existe um método de mineração de dados universal. Portanto, é possível que haja uma série de algoritmos cabíveis para uma mesma aplicação. Isso tudo depende do problema de negócio a ser solucionado e das características dos dados disponíveis para análise. Alguns dos principais critérios utilizados para a escolha da técnica mais adequada para atingir um determinado objetivo são: velocidade, facilidade de integração, habilidade para lidar com muitos atributos, facilidade de compreensão do modelo, facilidade de aplicação, utilidade e disponibilidade. Na maioria das vezes, a escolha do método exige a participação de pessoas que efetivamente entendem do assunto em questão, mesmo que não sejam especialistas na utilização e no manuseio computacional dos dados. Não existe uma técnica capaz de resolver todos os problemas de Mineração de Dados. Entretanto, cada técnica possui sua própria abordagem e seus critérios específicos que devem ser levados em consideração. Diferentes métodos servem para diferentes propósitos e cada um deles oferece suas vantagens e desvantagens. É recomendável que sejam testadas e combinadas diversas metodologias a fim de que comparações possam ser feitas e só então a melhor técnica (ou conjunto de técnicas) seja utilizada.

As técnicas de **lógica nebulosa**, também conhecida como lógica *fuzzy*, utilizam a teoria dos conjuntos nebulosos e são frequentemente utilizadas para capturar informações vagas que, em geral, são apresentadas em sua forma natural. Atualmente, ela é um elemento fundamental em diversos sistemas, sendo considerada uma técnica de excelência no universo computacional. Ela objetiva modelar o modo aproximado do raciocínio humano buscando desenvolver sistemas computacionais capazes de tomar decisões de forma racional em um ambiente muitas vezes incerto ou impreciso. Está baseada em palavras e não em números, ou seja, os valores de verdadeiro ou falso são expressados linguisticamente. Dessa maneira, ela pode ser considerada como um conjunto de princípios matemáticos para a representação do conhecimento baseado no grau de pertinência dos termos. Ela disponibiliza diversos métodos de manipulação de informações imprecisas, como os conceitos de pouco ou muito, alto ou baixo, quente ou frio, fornecendo assim uma resposta próxima a uma questão baseada em conhecimentos incompletos ou não exatos. Sistemas que se baseiam nesses procedimentos têm se mostrado de grande utilidade em uma variedade de operações de controle industrial e em tarefas de reconhecimento de padrões e de robótica.

A **classificação bayesiana** é uma técnica estatística baseada no teorema do matemático inglês Thomas Bayes que classifica um determinado elemento em uma determinada classe baseando-se na probabilidade do elemento pertencer a essa classe. Ela envolve probabilidade condicional e utiliza um dos mais importantes teoremas do estudo probabilístico atual. Através dela, é possível encontrar a probabilidade de um certo evento ocorrer, dada a probabilidade de um outro evento que já ocorreu. Ela altera estimativas iniciais com base em novas informações. Nela, são mostrados como fazer avaliações e como ajustá-las quando deparadas com novas informações. É uma abordagem analítica e interpretativa para o raciocínio probabilista e tem sido aplicada em inúmeras áreas, entre elas a estimação de risco operacional, a de projeto de jogos computacionais, inteligência artificial, imputação de dados, diagnóstico médico, entre outras. Frequentemente, é chamada de classificação simples ou ingênua, pois ela considera que o efeito do valor de um atributo sobre uma classe é independente dos valores dos outros atributos. O algoritmo calcula a probabilidade de que uma amostra desconhecida pertença a cada um dos grupos conhecidos. Ele obtém resultados melhores quando os valores coletados são discretos em vez de contínuos.

As **redes neurais artificiais** são tecnologias derivadas de pesquisas que utilizam a regressão generalizada e que tem origem na psicologia e na neurobiologia. São modelos inspirados na fisiologia do cérebro, na qual o conhecimento é fruto do mapa das conexões neuronais e dos pesos dessas conexões. Carvalho explica que esse tipo de técnica funciona “[...] sempre da camada de entrada para a de saída, ou seja, no sentido para frente, sendo apenas o cálculo de uma variável da regra de aprendizado que nos fornece a falsa impressão de que algo é retropropagado” (2001, p. 123). Uma rede neural pode ser vista como um conjunto de unidades de entrada e saída conectados por camadas intermediárias na qual cada ligação possui um peso associado. É composta de algoritmos e modelos computacionais inspirados pelo sistema nervoso central de um animal que são capazes de realizar o aprendizado da máquina, bem como o reconhecimento de padrões. Sua propriedade mais importante é a habilidade de evolução e de aprendizagem a partir de experiências e treinamentos, não ficando restritas a uma ordem sequencial pré-estabelecida. Algumas de suas

principais aplicações são: aproximação de funções, reconhecimento de padrões, classificações, sensoriamento remoto, robótica, processamento de voz e biometria.

Os **algoritmos genéticos** são um conjunto de operações matemáticas que procuram obter boas soluções para problemas complexos através de metodologias inspiradas pela biologia evolutiva, como a mutação, a hereditariedade e a seleção natural. Eles "[...] utilizam conceitos provenientes do princípio de seleção natural para abordar uma série ampla de problemas, em especial de otimização" (LUCAS, 2002, p.5). Consistem em aproximações computacionais de fenômenos vistos na natureza, como a mutação genética ou a reprodução sexuada. Eles funcionam mantendo uma população de estruturas, denominadas indivíduos ou cromossomos, operando sobre estas de forma semelhante à evolução das espécies. Essa população é inicializada aleatoriamente e evoluída através de operadores genéticos. O principal objetivo desse processo é o de realizar transformações na população fazendo com que, a cada nova geração, indivíduos cada vez mais evoluídos sejam desenvolvidos. Após o processo de cruzamento, é executado o operador de mutação, que tem por objetivo realizar modificações em determinadas propriedades genéticas de uma população de maneira aleatória. Esse processo continua até que população populações com indivíduos mais fortes e capazes sejam criadas ou que algum critério de parada tenha sido atingido.

A **indução por árvores de decisão** é uma técnica muito poderosa e amplamente utilizada quando se trata de problemas de classificação, estimativa, regressão e segmentação. Uma árvore de decisão, conforme explica Côrtes *et al,* é "[...] um fluxograma semelhante a uma estrutura de árvore, onde cada nó interno denota um teste em um atributo, cada ramo representa o resultado do teste e cada folha representa a distribuição dos registros" (2002, p.17). Consiste em uma hierarquização baseada nos estágios de decisão e na separação das classes e dos subconjuntos. São diagramas que permitem avaliar e representar problemas que envolvam decisões sequenciais, destacando os riscos e os resultados identificados nos diversos cursos de ação. A árvore sempre é lida de forma descendente, iniciando-se pelo nó raiz. Cada nível da árvore representa um estágio de decisão. Essas decisões resultam em regras para classificar conjuntos de dados. Uma de suas maiores vantagens é a de permitir a classificação de conjuntos com milhões de exemplos e centenas de atributos a uma velocidade razoável. Além disso, ela pode ser convertida para regras de classificação e, consequentemente, ser facilmente interpretadas e compreendidas a olho nu pelo ser humano. Normalmente, possuem um tempo de aprendizado relativamente rápido.

A **regressão linear** é uma técnica que se utiliza de uma equação para calcular o valor esperado de uma variável de interesse em função de outras variáveis, chamadas de variáveis independentes ou covariáveis. Consiste na obtenção de uma equação que tenta explicar a variação da variável dependente por meio da variação das variáveis independentes. Nesse método, a qualidade e a validade dos ajustes são avaliadas por meio de diagnósticos de regressão. A regressão linear múltipla envolve três ou mais variáveis, ou seja, uma variável dependente e pelo menos duas variáveis independentes. A **regressão não-linear**, por sua vez, consiste em uma análise observacional na qual os dados são modelados por uma função que é uma combinação não-linear de parâmetros do modelo e depende de uma ou mais variáveis independentes. Um modelo é considerado não-linear se pelo menos um dos seus parâmetros aparece no formato não-linear. Nesse método, além dos diagnósticos

tradicionais, diversos outros procedimentos devem ser seguidos. Através da transformação linear, alguns problemas de regressão não-linear podem ser convertidos em problemas de regressão linear. Em ambas as situações, os modelos utilizados são frequentemente ajustados utilizando a abordagem dos quadrados mínimos.

4. Funcionalidades e importância

A Mineração de Dados possui uma gama de aplicações excepcionalmente grande. Na área de serviço público, por exemplo, ela pode ser utilizada para descobrir necessidades públicas, melhorar o desempenho de algum serviço e auxiliar na tomada de decisões através de sistemas automatizados. Na área da saúde, a Mineração de Dados está se tornando cada vez mais essencial. Isso porque os dados dos registros de tratamento podem ser extraídos para explorar maneiras de reduzir custos e desenvolver melhores remédios. Tais dados podem também ser explorados para identificar e entender pacientes de alto custo e, através de dados oriundos de milhões de prescrições, operações e tratamentos, identificar padrões incomuns e descobrir fraudes. Na área industrial, ela pode ser de imensa ajuda em áreas como a de varejo, bancos e telecomunicações através das técnicas de classificação e agrupamento. Atualmente, ela é bastante utilizada para a detecção de comportamento de clientes, perfis de telespectadores, perfis de pacientes que têm doenças específicas, dentre outros, o que auxilia na criação de uma melhor imagem interna e externa da empesa.

Conforme explica Marino, “[...] as empresas devem ofertar produtos e serviços que atendam às necessidades e expectativas do mercado, que sejam úteis, que cubram custos, que garantam lucros e que tenham preços competitivos” (2006, p.1). Para atingir esses objetivos, é importante que a empresa, antes de mais nada, conheça o perfil de consumo dos seus clientes. Esse conhecimento precisa abranger vários níveis, desde o tipo de produto desejado até que tipo de ofertas estão dispostos a aceitar, mesmo que tais produtos não sejam essenciais no momento. Nesse contexto, a Mineração de Dados pode ser uma ferramenta extremamente útil, pois ela se utiliza dos dados das compras dos clientes no passado recente para traçar seus perfis de consumo. Setores como o de transporte, varejo e a internet são pioneiros na utilização da Mineração de Dados, pois lidam com imensas bases de dados, sendo que a competitividade nesses ramos é muito alta. Informações como idade, renda, profissão, estado civil, moradia própria ou alugada, bairro e cidade também devem ser levadas em consideração, pois permitem uma setorização ainda mais rigorosa dos clientes.

A Mineração de Dados é de extrema importância para organizações que possuam um banco de dados muito grande e necessitam tomar decisões de maneira ágil. Ela auxilia no desenvolvimento e na execução de estratégias, práticas e procedimentos para a gerência dos dados em uma organização. Frequentemente, muitas informações valiosas permanecem “escondidas” em meio a massas gigantescas de dados e, portanto, é praticamente impossível identifica-las através de linguagens de consulta tradicionais. Gerenciar esse turbilhão de dados e transformá-los em conhecimento de mercado pode ser uma atividade decisiva para a sobrevivência da empresa. Analistas humanos poderiam levar semanas ou até meses para correlacionar e descobrir alguma informação útil a olho nu. Através do processo de mineração, a empresa pode facilmente identificar as características comuns de clientes que compram

os mesmos produtos de uma empresa, predizer quais clientes provavelmente deixarão a empresa para um concorrente, revelar as diferenças entre um típico cliente de um mês em relação aos meses anteriores, entre outras atividades.

Devido à competição empresarial, clientes mudam de empresa com facilidade. Conforme explica Carvalho, em seu livro, “[...] detectar regularidades em fenômenos que ocorrem ao longo do tempo e poder prever tendências futuras é uma tarefa das mais importantes no mundo atual” (2001, p. 227). Através da Mineração de Dados, é possível verificar por que os clientes trocam uma empresa por outra e oferecer serviços, vantagens e ofertas que evitem essa fuga de clientes. É preferível fazer com que as empresas desenvolvam estratégias para manter os clientes em vez de gastar esforços na tentativa de adquirir outros. Dessa maneira, pode-se localizar quais produtos oferecer a quais clientes para fazer com que os mesmos permaneçam na empresa, ou mesmo localizar os clientes que podem sair da empresa sem representar prejuízo. Com dados de clientes, é possível definir hábitos de consumo e prever necessidades de outras classes sociais em outras cidades. É muito importante que as organizações conheçam o perfil dos seus clientes e ofereçam um atendimento personalizado de acordo com suas necessidades e preferências. Isso requer inteligência e aprendizado.

A funcionalidade de **identificação** tem por objetivo utilizar os padrões de dados para identificar a existência de um item, uma atividade ou um evento. Consiste na tarefa de detectar, discernir ou distinguir um elemento ou um conjunto de elementos que não se encaixam em um determinado padrão. Para que os padrões sejam descobertos, é necessário que os dados sejam simplificados a fim de que se possa descartar aquilo que é específico e focar naquilo que é genérico. Nesses casos, é recomendável que mecanismos de codificação sejam aplicados para reduzir o tamanho do conjunto de dados. Dados bem organizados e formatados colaboram para resultados eficientes durante a busca por conhecimento. Por exemplo, caso algum intruso tente romper um sistema, este pode ser detectado e identificado através dos programas executados, dos arquivos acessados e do tempo de CPU por sessão. Além disso, ela pode ser utilizada por empresas voltadas ao comércio através da identificação de perfis para determinados produtos com base nos dados, incluindo faixa etária, sexo, renda, entre outros, armazenados em banco. Assim, é possível conquistar um número maior de clientes e, consequentemente, impulsionar as vendas e a lucratividade da empresa.

A **classificação** é uma das funcionalidades mais comuns da Mineração de Dados e consiste na análise de dados em busca da definição de padrões que descrevem tendências futuras desses dados. De maneira resumida, é “[...] um modelo constituído de regras que permitem classificar os objetos de um banco de dados dentro de um número de classes pré-determinado” (NASCIMENTO; CRUZ, p. 2346). Consiste em tipificar, hierarquizar, categorizar, organizar ou nomear um dado ou um conjunto de dados por meio da identificação de propriedades comuns em seus atributos através de algoritmos supervisionados. Ela visa mapear um registro em uma base de dados a uma ou mais classes pré-definidas. Após esse processo, é possível aplicar esse método a novos registros que venham a ser incorporados na base de dados. Através dela, é possível, por exemplo, identificar em uma escola qual é a turma mais apropriada para um determinado aluno, atribuir palavras-chave a artigos jornalísticos, determinar quando uma transação de cartão de crédito pode ser uma fraude, identificar quando uma pessoa pode ser uma ameaça para a segurança de um sistema, entre outras

aplicações. Algumas das técnicas mais utilizadas nessa funcionalidade são: classificação bayesiana, árvores de decisão, redes neurais e algoritmo genético.

Um dos grandes desafios que as organizações enfrentam atualmente é o de se manterem competitivas à medida que as condições dos ambientes de negócios mudam e se tornam mais complexas. Assim sendo, é necessário que as empresas busquem otimizar seus recursos como uma tentativa de garantir sua perenidade no mercado, que está cada vez mais exigente e competitivo. Nesse contexto, uma das funcionalidades mais importantes da Mineração de Dados é a da **otimização**, que possui como objetivo minimizar o gasto de recursos limitados como tempo, dinheiro, espaço, ocupação de memória, matéria prima e mão-de-obra, ao mesmo tempo que busca maximizar variáveis de resultado como lucro, vendas, performance e produção. O objetivo do problema é normalmente expresso em uma função linear chamada de “função objetivo”. É fazer o uso de técnicas, modelos e métodos visando extrair o maior aproveitamento possível de qualquer atividade. É um ramo da matemática aplicada intimamente ligada à logística, pois até mesmo sistemas produtivos e industriais podem fazer uso de tais metodologias para alcançar um desempenho superior. Essa área “[…] estrutura processos, propondo um conjunto de alternativas de ação, fazendo a previsão e a comparação de valores, da deficiência e de custos” (LOESCH; HEIN, 2009, pág.1).

A funcionalidade de **estimação** consiste, basicamente, em determinar o valor mais provável de algum índice com base em dados do passado ou de dados de outros índices da mesma natureza. É a utilização de técnicas, modelos e métodos matemáticos para calcular, presumir ou estimar o valor de uma variável ou de um conjunto de variáveis. É o processo de predizer algum valor baseado num padrão já conhecido. Ela pode ser usada para executar uma tarefa de classificação partindo do princípio de que diferentes faixas de valores contínuos correspondem a diferentes classes ou grupos. Obviamente, tais estimativas estão sujeitas a erros, mesmo que pequenos. Entretanto, o objetivo é justamente determinar da melhor forma possível um valor baseando-se em outros valores de situações idênticas, mas nunca exatamente iguais. Ela pode ser utilizada, por exemplo, para estimar a pressão ideal de um ser humano baseando-se na sua idade, sexo, altura e massa corporal. Algumas das ferramentas mais comuns utilizadas nessa funcionalidade são: regressão linear, regressão múltipla, redes neurais artificiais, estatística e algoritmos genéticos.

A funcionalidade de **agrupamento** tem como objetivo formar grupos baseados no princípio de que esses grupos devem ser homogêneos em si e heterogêneos entre si. É a atividade de colecionar, aglomerar, amontoar ou organizar elementos de um banco de dados através da identificação de suas semelhanças ou atributos em comum. Ela pode, também, detectar a existência de diferentes grupos dentro de um determinado conjunto de dados e, caso isso ocorra, determinar esses grupos. Nessa aplicação, não existem classes predefinidas para classificar os registros em estudo. A base de dados sempre é de grande porte e cada elemento possui um número elevado de atributos ou características. Quando se deseja formar agrupamentos, um conjunto de atributos é selecionado e, em função da compatibilidade desses atributos, são formados grupos. Geralmente, é preciso realizar a transformação dos diferentes tipos de dados para uma escala comum, como [0.0, 1.0]. Em geral, as medidas de similaridade utilizadas são as medidas de distância tradicionais. Algumas de suas principais aplicações são: redes de

comunicações, computação médica, biologia computacional, redes de computadores, computação visual e gráfica, entre outras.

A **análise de desvios** é uma funcionalidade cujo principal objetivo é detectar mudanças de comportamentos através da comparação entre determinadas ações e padrões pré-estabelecidos. Por exemplo, ao avaliar o comportamento de compra de um usuário de um determinado cartão de crédito, é possível identificar alguma alteração brusca de comportamento e imediatamente verificar se houve ou não uma fraude no cartão de crédito. O mesmo vale para clientes em um plano de saúde ou qualquer outro usuário de um sistema automatizado cujos dados são armazenados em banco. Uma funcionalidade muito semelhante a essa é a de **análise de evolução**, que observa, estuda e descreve a regularidade de tendências ou modelos para objetos cujo comportamento sofre alterações ao longo do tempo. Ela modela tendências através da classificação, comparação, caracterização ou agrupamento de dados relacionados com o tempo. Por exemplo, a ampliação de um estoque, de uma criação de animais ou de uma plantação. Isso requer a utilização de diversas ferramentas estatísticas, incluindo cálculos das médias, medianas e dos desvios-padrão, e a análise da distribuição dos dados através de gráficos como o de ogiva, histograma ou *box-plot*.

5. Considerações finais

Conforme abordado ao longo do artigo, a Mineração de Dados é uma ferramenta de apoio com papel fundamental na gestão de informação dentro das organizações. Ela se apresenta confiável, ágil, eficiente e capaz de capturar informações valiosas, as quais podem servir de apoio nas tomadas de decisão. Muitos empreendimentos já coletam quantidades grandes de informações de diversas fontes; entretanto, para adquirir conhecimento, é preciso fazer correlações e identificar padrões entre elas. Em virtude do imenso volume de dados coletados diariamente e da altíssima competitividade entre as organizações, torna-se inviável a análise dessas informações da maneira tradicional. Ela tem a vantagem de encontrar padrões de relacionamento em meio a turbilhões de dados sem a tendenciosidade e a limitação de uma análise baseada apenas na intuição humana. Ela interpreta os dados a fim de produzir uma quantidade de padrões úteis, válidos e de fácil entendimento. As experiências e pesquisas nessa área têm proporcionado resultados satisfatórios para as empresas que a adotam. Através do conhecimento adquirido, empresas podem tornar ágil a tomada de decisão, minimizar custos, melhorar o relacionamento com o cliente, aumentar receitas e, consequentemente, garantir a sua perenidade no mercado.

Atualmente, existem dezenas de ferramentas comerciais para a atividade de mineração. Entretanto, é curioso saber que a popularidade de tais *softwares* é relativamente baixa se comparada à popularidade das ferramentas de *Data Warehouse* e OLAP, por exemplo. Grande parte dessas ferramentas exige que o usuário tenha algum conhecimento prévio acerca do funcionamento dos algoritmos de mineração implementados no sistema. Felizmente, o mercado demonstra estar realizando esforços na tentativa de corrigir esses problemas. Empresas como a Oracle e a Microsoft já estão desenvolvendo recursos inovadores para facilitar a utilização desses algoritmos. Assim, dois problemas podem ser resolvidos de uma só vez: o custo alto das ferramentas e a impossibilidade de mineração direta. Tais técnicas precisam ser

desenvolvidas ou adaptadas para tarefas e domínios específicos. Na Engenharia de *Software*, por exemplo, os dados obtidos podem resultar na descoberta de padrões capazes de melhorar a qualidade e a produtividade dos softwares desenvolvidos. Considerando o volume de dados, muitas vezes em proporções assustadoras, e a demanda por alto desempenho, presume-se que as novas técnicas escaláveis de mineração serão de imensa contribuição para a evolução e o desenvolvimento da área.

Referências

CARVALHO, Luiz Alfredo Vidal de. **Datamining**: a mineração de dados no marketing, medicina, economia, engenharia e administração. São Paulo: Editora Érica LTDA, 2001.

LOESCH, Claudio; HEIN, Nelson. **Pesquisa operacional:** fundamentos e modelos. São Paulo: Saraiva, 2009.

MARINO, Lúcia Helena Fazzane de Castro. **Gestão da qualidade e gestão do conhecimento:** fatores-chave para produtividade e competitividade empresarial. Disponível em: http://www.simpep.feb.unesp.br/anais/anais_13/artigos/598.pdf. Acesso em: 28 de maio de 2017.

LUCAS, Diogo C. **Algoritmos Genéticos:** uma Introdução. Disponível em: http://www.inf.ufrgs.br/~alvares/INF01048IA/ApostilaAlgoritmosGeneticos.pdf. Acesso em: 29 de maio de 2017.

CÔRTEZ, Sérgio da Costa; PORCARO, Rosa Maria; LIFSCHITZ, Sérgio. **Mineração de Dados** – Funcionalidades, Técnicas e Abordagens. Disponível em: http://www.dbd.puc-rio.br/depto_informatica/02_10_cortes.pdf. Acesso em: 31 de maio de 2017.

NASCIMENTO, Lídice Cabal; CRUZ, Carla Bernardete Madureira. **Mineração de dados e adaptação de modelos de classificação de cobertura e uso da terra para imagem Worldview.** Disponível em: http://www.dsr.inpe.br/sbsr2013/files/p0789.pdf. Acesso em: 04 de junho de 2017.